(12)

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT #63	2. GOVT ACCESSION NO.
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Three dimensional object recognition using an unsupervised neural network: Understanding the distinguishing features.	5. TYPE OF REPORT & PERIOD COVERED Technical Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) N. Intrator, J.I. Gold, H.H. Bulthoff, and S. Edelman	8. CONTRACT OR GRANT NUMBER(s) N00014-91-J-1316
9. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Brain and Neural Systems · Brown University Providence, Rhode Island 02912	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS N-201-484
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel & Training Research Program Office of Naval Research, Code 442PT Arlington, Virginia 22217	12. REPORT DATE December 23, 1992
	13. NUMBER OF PAGES 7 pages
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Publication in part or in whole is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Published in Y. Feldman and A. Bruckstein, editors, Proceeding of the 8th Israeli Conference on AICV, pages 113-123. Elsevier, 1991.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Feature Extraction
BCM Theory
Object Recognition
Unsupervised Learning
Projection Pursuit

93 2 2 004 4/29/93

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A novel method for feature extraction has been applied to a problem of three dimensional object recognition (Intrator and Gold, 1991). The method is related to recent statistical theory (Huber, 1985; Friedman, 1987) and is derived from a biologically motivated computational theory (Bienenstock et al., 1982). Results of an initial study replicating recent psychophysical experiments (Bulthoff and Edelman, 1991) demonstrated the utility of the proposed method for feature extraction. We describe further experiments designed to analyze the nature of the extracted features, and their relevance

to the theory and psychophysics of object recognition.

Three-Dimensional Object Recognition Using an Unsupervised Neural Network: Understanding the Distinguishing Features

Nathan Intrator*
Center for Neural Science
Brown University
Providence, RI 02912, USA

Josh I. Gold
Center for Neural Science,
Brown University,
Providence, RI 02912, USA

Heinrich H. Bülthoff
Dept. of Cognitive and Linguistic Sciences,
Brown University,
Providence, RI 02912, USA

Shimon Edelman
Dept. of Applied Mathematics
and Computer Science,
Weizmann Institute of Science,
Rehovot 76100, Israel

Abstract

A novel method for feature extraction has been applied to a problem of three-dimensional object recognition (Intrator and Gold, 1991). The method is related to recent statistical theory (Huber, 1985; Friedman, 1987) and is derived from a biologically motivated computational theory (Bienenstock et al., 1982). Results of an initial study replicating recent psychophysical experiments (Bülthoff and Edelman, 1991) demonstrated the utility of the proposed method for feature extraction. We describe further experiments designed to analyze the nature of the extracted features, and their relevance to the theory and psychophysics of object recognition.

1 Introduction

Object recognition may be accomplished via a comparison between an image and a set of templates that represent known objects. However, since the number of different objects that are to be recognized — including possible transformations of each object — can be very large, approaches more sophisticated than simple template matching are required. One possibility is to represent objects by low-dimensional sets of features. What such features could be is, however, not at all clear, and is subject to current research (see review in Edelman, 1991).

Intrator (1990) proposed a feature extraction method that is related to recent statistical theory (Huber, 1985; Friedman, 1987), and is based on a biologically motivated model of neuronal plasticity (Bienenstock et al., 1982). This led to a model for object recognition (Intrator and Gold, 1991) which was evaluated by simulating psychophysical experiments of 3D object recognition (see Bülthoff and Edelman, 1991). The model's ability to generalize recognition to novel views compared favorably to the psychophysical results. The success of the model has led to an in-depth study of the nature of the features extracted for recognition. We start

*Research was supported by the National Science Foundation, the Army Research Office, and the Office of Naval Research.

with a brief overview of this recognition model, focusing on feature extraction in a statistical framework, and review both the experimental paradigm and our results from a previous study. We then describe the current study examining the effects of occluding these features in the images.

2 What Are Features of Recognition

The discussion of the issue of features of recognition in recent psychological literature is relatively scarce (LaBerge, 1976). A possible reason for that may be the predominance of structural models of recognition, of which a recent example is the Recognition By Components (RBC) theory (Biederman, 1987). Structural models, which have supplanted previously widespread theories based on invariant feature spaces, represent objects in terms of a small set of generic parts and spatial relations among parts. Naturally, the question of possible existence and relevance of a variety of features, as well as the dimensionality reduction problem, does not arise in the structural approach (see, however, Edelman 1991).

In comparison, invariant feature theories follow the standard approach of statistical pattern recognition in postulating that objects are represented by clusters of points in multidimensional feature spaces (Duda and Hart, 1973). Although some attempts have been made to generate and verify specific psychophysical predictions based on the feature space approach (see especially (Shepard, 1987)), feature-based psychological models of recognition do not seem to be computationally adequate to allow the inference of their stand on the issue of dimensionality reduction and feature learning.

Results from recent psychophysical experiments (Edelman and Bülthoff, 1990; Edelman et al., 1991), namely, the improvement in performance with increasing stimulus familiarity, are compatible with a feature-based recognition model which extracts problem-specific features in addition to universal ones. Specifically, the subject's ability to discern key elements of the solution appears to increase as the problem becomes more familiar. This finding suggests that some of the features used by the visual system are based on the task-specific data, and therefore raises the question of how can such features be extracted.

The model proposed by Intrator and Gold (1991) which is briefly described below puts the emphasis on the dimensionality reduction; namely, it seeks features of a set of objects that would best distinguish among the members of the set. This method does not rely on a general predefined set of features. This is not to imply, however, that features extracted by this method are useful only in recognition of the original set of images from which the features were extracted. In fact, the potential importance of this set of features is related to their invariance properties, or their ability to generalize. Invariance properties of this feature extraction method have already been demonstrated in speech recognition, in which the extracted features had better generalization properties across speakers and across phonemes than features found by other dimensionality reduction methods such as back-propagation and principal components analysis (Intrator, 1990; Intrator and Tajchman, 1991).

2.1 Feature Extraction in High Dimensional Space – the BCM Model

From a mathematical viewpoint, extracting features from gray level images is related to dimensionality reduction in a high dimensional vector space, in which an $n \times k$ pixel image is considered to be a vector of length $n \times k$. In such high dimensional spaces the *curse of dimensionality* (Bellman, 1961) says that it is impossible to base the recognition on the high dimensional vectors, because the number of patterns needed to train a classifier increases exponentially with

the dimensionality. Therefore, dimensionality reduction should take place before classification is attempted. Due to the large number of parameters involved, a feature extraction method that uses the class labels of the data may be biased to the training data, resulting in features with poor generalization or invariance properties. Thus, feature extraction should be at least partially unsupervised.

The best-known method for extracting features is principal component analysis. It has been argued, however, that principal component features may not retain the structure needed for classification (Duda and Hart, 1973; Huber, 1985). A more general and powerful method for feature extraction is Projection Pursuit, and its unsupervised version, Exploratory Projection Pursuit (Friedman and Tukey, 1974; Friedman, 1987). This method has been extended in various directions, and is reviewed in (Huber, 1985). The idea behind projection pursuit is to pick *interesting* low dimensional projections of a high dimensional "point cloud", by maximizing an objective function called projection index.

For the purpose of pattern classification, it is important to concentrate on dimensionality reduction methods that allow discrimination between classes, rather than faithful representation of the data. This leaves out methods such as factor analysis (Harman, 1967, for review) which tend to combine features with high correlation.

Various objective functions are motivated by different assumptions about the notion of what constitutes an *interesting* feature in a data set. In the first approximation, one may consider only features defined by linear (or semi-linear) projections of high dimensional data. A statement recently formulated by Diaconis and Freedman (1984) says that for most high-dimensional data "clouds", most low-dimensional projections are approximately normal. This finding suggests that the important information in the data is conveyed in those directions whose single dimensional projected distribution is far from Gaussian. Friedman (1987), and Hall (1989) define interesting projections by measuring directly deviation from normality. Motivated by the fact that high dimensional clusters translate to low dimensional multi-modal projected distributions, Intrator (1990) presented a multiple feature extraction method that seeks multimodality in the projections. This method is based on a modified version of the BCM neuron (Bienenstock, Cooper and Munro, 1982), extended to a non-linear neuron model for reducing sensitivity to outliers. The lateral inhibition network architecture and the simplicity of the projection index makes this method computationally practical for simultaneous extraction of several interacting features from high dimensional spaces. The biological relevance of the theory has been extensively studied (Saul and Daniels, 1986; Bear et al., 1987; Bear and Cooper, 1988) and it was shown that the theory is in agreement with several classical visual deprivation experiments (Clothiaux et al., 1991).

The unsupervised feature extraction/classification method used in the present study is illustrated in Figure 1. Various other approaches call for dimensionality reduction prior to classification, e.g., using the RCE network (Reilly et al., 1982) as a classifier and back-propagation network for dimensionality reduction (Reilly et al., 1987; Zemani et al., 1989), or using the unsupervised charge clustering network (Cooper and Scofield, 1988). We note that the classifier performing classification on the extracted features may affect the generalization properties of the entire scheme. For example, using features extracted by a BCM network, a back-propagation classifier performed better than a k-nearest neighbor classifier (k-NN) in the speech recognition experiments (Intrator, 1991). Moreover, since the features are extracted using a projection index that favors multimodality, the resulting projections may be close to a mixture of Gaussians. This suggests that performing the classification with a GRBF-type network (Moody and Darken, 1988; Poggio and Girosi, 1990) may be more appropriate. However, in this paper our main concern is with the properties of the extracted features, not with classification (which

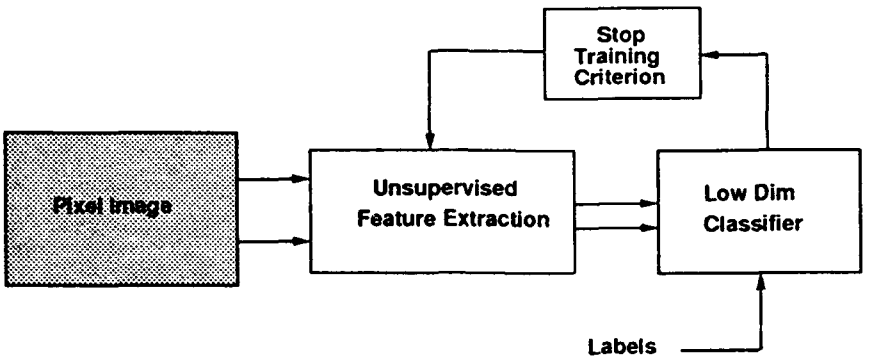

Figure 1: Low dimensional classifier is trained on the features extracted from the high dimensional data. Training of the feature extraction network stops, when misclassification rate drops below a predetermined threshold on either the same training data (cross validatory test) or on a different testing data.

was, therefore, implemented by a simple k-NN classifier).

2.2 Experimental paradigm

Previous work in the study of object recognition has led to the development of an experimental paradigm (Bülthoff and Edelman, 1991) designed to test generalization from familiar to novel views of three dimensional objects. The paradigm is useful for the present study because it can be applied both to human subjects and to computer models.

We have used as stimuli novel, wire-like objects, developed by Edelman and Bülthoff (1990, 1991) . These objects proved to be easily manipulated, and yet complex enough to yield interesting results. Wires were also used in an effort to simplify the problem for the feature-extractor, as they provided little or no occlusion of the key features from any viewpoint. Wire objects were generated by the Symbolics S-Geometry^{TM} modelling package, and rendered with a visualization graphics tool (AVS, Stardent, Inc.). Each object consisted of seven connected equal length segments, pointing in random directions and distributed equally around the origin (for further details, see Edelman and Bülthoff, 1990).

Each experiment consisted of two phases, training and testing. In the training phase subjects were shown the target object from two standard views, located 75 degrees apart along the equator of the viewing sphere. The target oscillated around each of the two standard orientations with an amplitude of ± 15 degrees about a fixed vertical axis, with views spaced at 3-degree increments (see Figure 2). Test views were located either along the equator – on the minor arc bounded by the two standard views (INTER condition) or on the corresponding major arc (EXTRA condition) – or on the meridian passing through one of the standard views (ORTHO condition). Testing was conducted according to a two-alternative forced choice (2AFC) paradigm, in which subjects were asked to indicate whether the displayed image constituted a view of the target object shown during the preceding training session. Test images were either unfamiliar views of the training object, or random views of a distractor (one of a distinct set of objects generated by the same procedure).

To apply the above paradigm to the BCM network, the objects were imported into a 3D visualization package (AVS, Stardent Inc.). All objects were displayed in a 63x63 array, under simulated illumination that combined ambient lighting of relative strength 0.3 with a point source of strength 1.0 at infinity. The raw image "seen" by the network consisted of an array

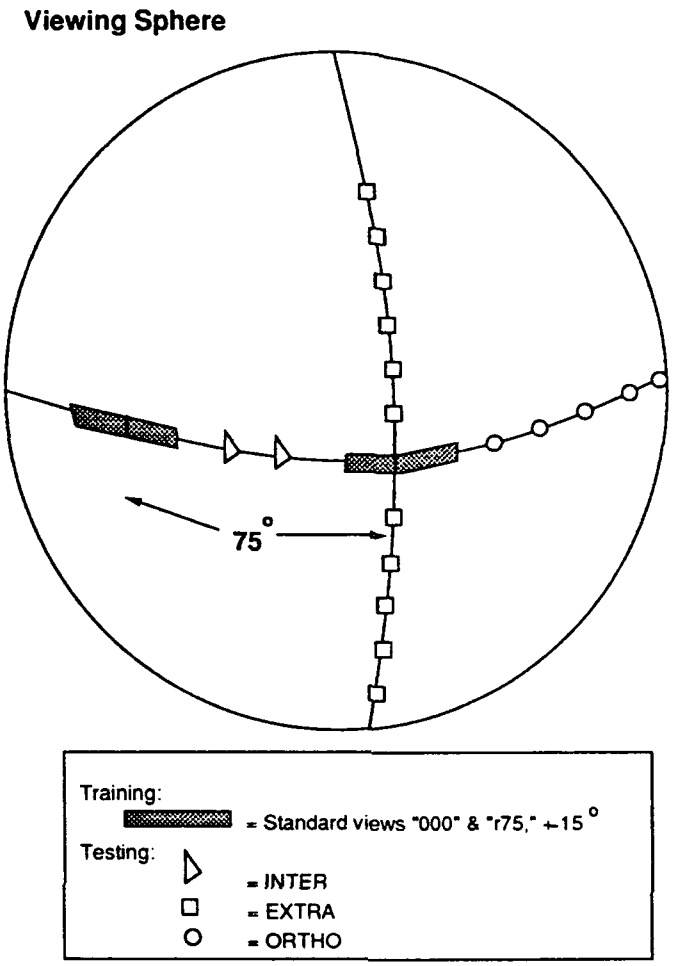

Figure 2: The view sphere visualization of the experimental paradigm.

of gray-scale values ranging from 0 to 255. The study described below involved six-way classification, which is more difficult than the 2AFC task used in the psychophysical experiments.

2.3 Results of the Previous Study

The six wires used in the experiments are depicted in Figure 3. Results of the previous study (Intrator and Gold, 1991) demonstrated that the BCM network could in fact extract rotation-invariant features which were useful in solving this 3D object recognition problem. In particular, two results from the psychophysical studies were replicated by the BCM network: (1) error rates increased steadily with misorientation relative to the training view; (2) generalization in the horizontal direction was better than in the vertical direction. Table 1 summarizes the performance of human subjects and the predictions of several computational theories in relation to these two points.

Given the task of recognizing the six wires, the network extracted features that corresponded to small patches of the different images, namely areas that either remained relatively invariant under the rotation performed during training, or represented distinctive features of specific wires. The classification results are in good agreement with the psychophysical data: (1) the error rate was the lowest in the INTER condition, (2) recognition deteriorated to chance level

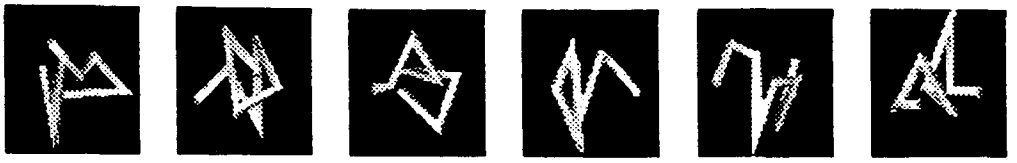

Figure 3: The six wires used in the computational experiments, as seen from a single view point.

THEORY	Ability to generalize from familiar to unfamiliar views	Generalization ability for unfamiliar views across horizontal vs. vertical training direction
RAL	uniformly good	same
LC	uniformly good for INTER and EXTRA	same
GRBF	steady decrease w/ increasing unfamiliarity	better for horizontal
BCM	steady decrease w/ increasing unfamiliarity	better for horizontal
Human	steady decrease w/ increasing unfamiliarity	better for horizontal

Table 1: Schematic comparison of several models for object recognition; RAL – Recognition by Alignment (Ullman and Basri, 1991), LC – Linear Combination (Basri and Ullman, 1989), GRBF – Generalized Radial Basis Functions (Poggio and Girosi, 1989).

with increased misorientation in the EXTRA and ORTHO conditions, and (3) horizontal training led to a stronger performance in the INTER condition than did vertical training. The first two points were interpreted as resulting from the ability of the BCM network to extract rotation invariant features. Indeed, features which exist on all training views would be expected to correspond to the INTER conditions. EXTRA and ORTHO views, on the other hand, are less familiar and therefore yield worse performance, and also may require features other than the rotation-invariant ones extracted by the model. The horizontal-vertical asymmetry (the third point mentioned above) might be related to an asymmetric visual field in humans (Hughes, 1977). Consequently, this asymmetry was modeled by increasing the resolution along the horizontal axis. Specifically, the aspect ratio between horizontal and vertical acuity was set to 2.00 for horizontal direction training, while for vertical training the aspect ratio was 0.50.

3 Examining the Features of Recognition

To understand the meaning of the features extracted by the BCM network under the various conditions and to establish a basis for further comparison between the psychophysical experiments and computational models, we developed a method for occluding key features from the images and examining the subsequent effects on the various recognition tasks.

3.1 The Occlusion Experiment

For this set of experiments, the procedure described above was modified so that some of the features previously extracted by the network could be occluded in the images during training and/or testing. Each input to a BCM neuron in our model corresponds to a particular point on a 2D input image, while "features" correspond to combinations of excitatory and inhibitory inputs. Assuming that inputs with strong positive weights constitute a significant proportion of the features, we select inputs whose weights exceed a preset threshold from an previously-trained synaptic weight matrix and occlude (i.e., set to black) the corresponding pixels in the input image.

The first hypothesis we tested concerns the general "usefulness" of the extracted features for recognition. If the features extracted by the BCM network do capture rotation-invariant aspects of the object and can support recognition across a variety of rotations, then occluding those features during training should lead to a pronounced and general decline in recognition performance of the model. In particular, recognition should deteriorate most significantly in the INTER and EXTRA cases, since they lie along the direction of rotation during training and therefore can be expected to rely to a larger extent on rotation-invariant features. Little change should be seen in the ORTHO condition, on the other hand, because recognition of ORTHO views, which are situated outside the direction of rotation defined by the training phase, does not benefit from rotation invariant features.

Figure 4 shows a synaptic weight matrix generated in the previous study, and the set of wires with the corresponding features occluded. Also shown is the control case with randomly occluded pixels in the input images.

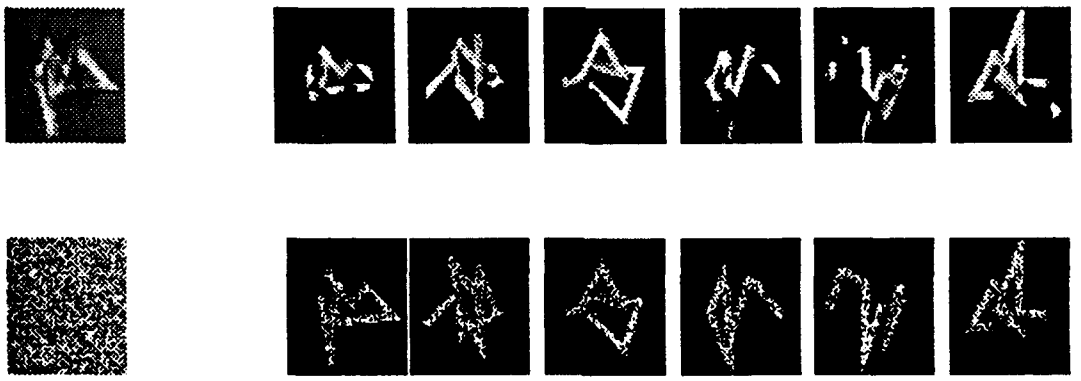

Figure 4: Wires at the top occluded with features taken from a trained matrix (top left). Wires at the bottom occluded with features taken from a randomized weight matrix (bottom left).

3.2 Results

Figures 6–9 summarize the results of the experiments. The first two show results from the previous study, which replicated corresponding results of psychophysical experiments (namely, the strong INTER performance, and the weaker performance under EXTRA and ORTHO conditions). The better generalization to novel views within the horizontal direction as compared to the vertical direction was also replicated.